Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 5529389, 9 pages
https://doi.org/10.1155/2021/5529389

Research Article

Using Recursive Feature Selection with Random Forest to Improve Protein Structural Class Prediction for Low-Similarity Sequences

Yaoxin Wang,[1] Yingjie Xu,[2] Zhenyu Yang,[1] Xiaoqing Liu,[3] and Qi Dai[1]

[1]*College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[2]*Qixin School, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[3]*College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, China*

Correspondence should be addressed to Qi Dai; daiailiu04@yahoo.com

Received 7 February 2021; Accepted 28 April 2021; Published 8 May 2021

Academic Editor: Lin Lu

Many combinations of protein features are used to improve protein structural class prediction, but the information redundancy is often ignored. In order to select the important features with strong classification ability, we proposed a recursive feature selection with random forest to improve protein structural class prediction. We evaluated the proposed method with four experiments and compared it with the available competing prediction methods. The results indicate that the proposed feature selection method effectively improves the efficiency of protein structural class prediction. Only less than 5% features are used, but the prediction accuracy is improved by 4.6-13.3%. We further compared different protein features and found that the predicted secondary structural features achieve the best performance. This understanding can be used to design more powerful prediction methods for the protein structural class.

1. Introduction

Protein structural class is the basic research field in protein research and makes a significant contribution to the research on protein function, protein folding rate, DNA binding site, and protein folding recognition, as well as reducing the search of conformational space and realizing the prediction of the tertiary structure [1–7]. In recent years, the gap between protein sequences and protein structures is becoming larger and larger with the development of sequencing technology, and it is relatively slow to identify three-dimensional structures by experimental methods. Therefore, it is necessary to develop computational methods for fast and accurate determination of protein structural classes.

The protein structures are determined by their sequences. Therefore, protein structure classes can be directly determined based on the sequence information, which can further guide biological experiments and reduce experimental costs. Many protein structural class prediction methods have been proposed since the concept of the protein structure class was put forward [3–5, 7–11]. At first, protein structural class prediction is designed based on the protein composition [1, 12, 13], such as short peptide composition [14–16], pseudo amino acid composition [17–20], and functional domain composition collocation [21]. Amino acid composition (AAC) is calculated according to the ratio of 20 amino acid residues in the sequence and denoted as a numerical vector as the sequence characteristic information [14–16]. However, it did not take the interaction and physicochemical properties of amino acids into account. Pseudo amino acid composition (PseACC) was further proposed as the characteristic information of protein [17–22], which does not merely consider the amino acid residues' composition but also considers the physical and chemical properties such as hydrophobicity of amino acid residues. In addition, the characteristic information is extracted by calculating the peptide components [23], which takes into account the sequence factors among amino acid residues.

The prediction method based on sequence-based features performs well on the high similarity data set, while their precision on the low-similarity data set is only about 50%. Some improved feature extraction methods need to be put forward

urgently. Kurgan et al. introduced a SCPRED method with the help of the predicted secondary structures [24]. Zhang et al. calculated a TPM matrix to represent the prediction of secondary structural features [25]. Dai et al. also proposed a statistical feature of the secondary structural features for protein structural class prediction [26]. Ding et al. constructed a multidimensional representing vector as the predicted secondary structure features, and some methods of fuse multiple features are also designed [27]. Chen et al. proposed a multifeature fusion method that combines structural information with physical chemistry [28, 29]. Nanni et al. introduced a prediction method that combines the characteristics of the first-level sequence and the characteristics of the second-level structure [30]. Wang et al. have combined improved simplified PSSM with secondary structural features for protein structural class prediction [31].

With help of the above features, prediction accuracy was improved over 80% for several low-similarity benchmark data sets, but some problems still exist in their development. In order to improve the efficiency of the prediction models, some research integrated different protein features to establish a prediction model. However, it is worth noting that the simple combination of the different features does not necessarily improve the prediction performance. If the combination is not appropriate, it may even offset the information contained in each other, which not only causes the redundancy of information but also increases the complexity and computation of the model.

With the above problems in mind, we proposed a scheme to predict the protein structural classes using the recursive feature selection with random forest. We first explored protein content features, protein position features, reduced combined features, and predicted secondary structural features and discussed their contribution for protein structural class prediction. We then proposed a recursive feature selection method to select important features from the above feature set, where the relative importance index of each feature is calculated based on the random forest algorithm. At last, the features are selected according to their relative importance value. Through a comprehensive comparison and discussion, some novel valuable guidelines for use of the recursive feature selection and protein features are obtained.

2. Materials and Methods

2.1. Data Sets. Four widely used low-similarity benchmark data sets are selected for comparison with existing methods [24, 25, 32–37]. The first data set is 25PDB, with sequence homology of 25%, which was originally published in [32, 33]. It contains 1673 proteins and domains, which are downloaded from PDB and scanned with high resolution. The second data set is D640, which has 25% sequence identity. It is composed of 640 proteins, and the classification tags are from the SCOP database [32, 33]. The third data set is FC699, in which 858 sequences have 40% low identity. The last data set, denoted as 1189, has 40% sequence identity. It is composed of three-dimensional structure data of 1092 proteins, which are downloaded from the RCSB protein database, and PDB ID is listed in [38]. Table 1 provides more detailed information about these low-similarity benchmark data sets.

Table 1: Protein distribution of different structural classes among four protein data sets.

Data set	All-α	All-β	α/β	$\alpha+\beta$	Total
25PDB	443	443	346	441	1673
D640	138	154	177	171	640
FC699	130	269	377	82	858
1189	223	294	334	241	1092

2.2. Sequence Content Feature. There are a large number of statistical literatures, in which a sequence is interpreted as a series of symbols. A k-word is a sequence of k-consecutive letters in a sequence. For the sequence s with length m, the count of k-word w, represented by $c(w)$, is the number of times w appears in the sequence s. Here, the k-word is allowed to overlap in the sequence. The sequence content can be described by the frequencies of the k-word, and it can be represented by an n-dimensional vector C_k^s:

$$C_k^s = (c(w_{k,1}), c(w_{k,2}), \cdots, c(w_{k,n})), \tag{1}$$

where n is the total number of all possible k -words. Then, the sequence content features can be calculated as

$$SCF_k^s = \left(\frac{c(w_{k,1})}{m-k+1}, \frac{c(w_{k,2})}{m-k+1}, \cdots, \frac{c(w_{k,n})}{m-k+1}\right). \tag{2}$$

This work calculates SCF_1^s and SCF_2^s to construct the sequence content features.

2.3. Sequence Position Feature. In addition to the sequence content features, we also pay attention to position distribution of these k-word elements. Given a k-word, we first transformed a protein structural sequence into several position signal sequences. If the interval distance $\mathrm{Dis}(w_{k,i})$ of the given k-word $w_{k,i}$ is equal to 1, the consecutive k-word $w_{k,i}$ will form a structure and motif domain. Otherwise, they belong to two different domains. Given the $\mathrm{Dis}(w_{k,i})$ and the integer t, we calculate the probability that $\mathrm{Dis}(w_{k,i})$ takes the value t, and the probability distribution of the $\mathrm{Dis}(w_{k,i})$ will be obtained. The numerical characteristics semimean $\text{Semi-E}_k(w)$ and semivariance $\text{Semi-D}_k(w)$ are defined by

$$\begin{aligned}
\text{Semi-E}_k(w) &= \sum_{\mathrm{Dis}(w_k)=1}^{t} \mathrm{Dis}(w_k) \times \mathrm{P}(\mathrm{Dis}(w_k)),\\
\text{Semi-D}_k(w) &= \sum_{\mathrm{Dis}(w_k)=1}^{t} \left((\mathrm{Dis}(w_k))^2 \times \mathrm{P}(\mathrm{Dis}(w_k))\right)\\
&\quad - \left[\sum_{\mathrm{Dis}(w_k)=1}^{t} \mathrm{Dis}(w_k) \times \mathrm{P}(\mathrm{Dis}(w_k))\right]^2.
\end{aligned} \tag{3}$$

The sequence position feature of the standard Semi-D_k to Semi-E_k is defined as

$$\mathrm{SPF}_k(w) = \frac{\text{Semi-E}_k(w)}{\text{Semi-D}_k(w)}. \quad (4)$$

$\mathrm{SPF}_k(w)$ is the variability of the k-word w in relation to its population mean [26], and we calculate $\mathrm{SPF}_1(w)$ and $\mathrm{SPF}_2(w)$ to construct the sequence position features.

2.4. Reduced Sequence Feature. Hydrophilicity is an important physical and chemical property of amino acids. According to the hydrophilicity of amino acids, 20 kinds of amino acids can be divided into three categories: internal group, external group, and ambivalent group. The reduction of protein sequences is defined according to the following rule:

$$F(S(i)) = \begin{cases} I, & \text{if } S(i) = F, I, L, M, V, \\ E, & \text{if } S(i) = D, E, H, K, N, Q, R, \\ A, & \text{if } S(i) = D, E, H, K, N, Q, R, \end{cases} \quad (5)$$

where $S(i)$ represents the i-th letter in protein sequence s and $F(S(i))$ represents the substitution for $S(i)$.

With help of the $F(S(i))$, a protein sequence can be transformed into a reduced sequence, which contains only three letters I, E, and A. For example, given a protein sequence $S = ESHFTCISLNEYAMQ$, we can get its reduced protein sequence $F(S) = EAEIAAIAIEEAAIE$. Here, we calculate the sequence composition and position features of the reduced sequence to combine reduced sequence features.

2.5. Predicted Secondary Structural Features. The protein sequence feature achieves promising results in the protein structural class prediction, but its accuracy is limited. Some studies have shown that the content and spatial arrangement of secondary structural elements are also important factors affecting the complex function or structure of proteins. Therefore, one of the methods to improve the prediction accuracy is to add secondary structural features to the feature set [24–31]. In this work, PSI-PRED is used to predict the secondary structure sequence [39], and the 11 widely used predicted secondary structural features are calculated to improve protein structural class prediction [40].

(1) Predicted secondary structure element content ($\mathrm{content}_{\mathrm{SE}}$): given a predicted secondary structure, the content of its predicted secondary structure elements $\mathrm{content}_{\mathrm{SE}}$ can be calculated by the following formula

$$\mathrm{content}_{\mathrm{SE}} = \frac{\mathrm{Count}_{\mathrm{SE}}}{\sum_{x \epsilon \{C,H,E\}} \mathrm{Count}_x}. \quad (6)$$

H, E, and C denote α-helix, β-strand, and coil, respectively.

(2) First- and second-order composition moment vector (CMV), another important structure feature, can be calculated as follows:

$$\mathrm{CMV}_{\mathrm{SE}}^k = \frac{\sum_{j=1}^{\mathrm{Count}_{\mathrm{SE}}} \mathrm{PO}_{\mathrm{SE}_j}^k}{\prod_{d=1}^{k} (N-d)}, \quad (7)$$

where $\mathrm{PO}_{\mathrm{SE}_j}^k$ denotes the secondary structure element at the j-th position in the secondary structure sequence with length N, and k is the vector order.

(3) Length of the longest segment ($\mathrm{MaxSeg}_{\mathrm{SE}}$):

$$\mathrm{MaxSeg}_{\mathrm{SE}} = \mathrm{MaxLen}(\mathrm{SEG} : \mathrm{SEG}_{\mathrm{SE}}), \quad (8)$$

where MaxLen denotes the maximal segment length function and $\mathrm{SEG}_{\mathrm{SE}}$ is the segments that consist of the structure element SE.

(4) Normalized length of the longest segment ($\mathrm{NMaxSeg}_{\mathrm{SE}}$):

$$\mathrm{NMaxSeg}_{\mathrm{SE}} = \frac{\mathrm{MaxLen}(\mathrm{SEG} : \mathrm{SEG}_{\mathrm{SE}})}{N}, \quad (9)$$

where N is the sequence length.

(5) Average length of the segment ($\mathrm{AvgSeg}_{\mathrm{SE}}$):

$$\mathrm{AvgSeg}_{\mathrm{SE}} = \frac{\sum \mathrm{Len}(\mathrm{SEG} : \mathrm{SEG}_{\mathrm{SE}})}{\mathrm{Content}_{\mathrm{SEG}_{\mathrm{SE}}}}, \quad (10)$$

where Len is the segment length function and $\mathrm{Content}_{\mathrm{SEG}_{\mathrm{SE}}}$ denotes the content of the $\mathrm{SEG}_{\mathrm{SE}}$.

(6) Normalized average length of the segment ($\mathrm{NAvgSeg}_{\mathrm{SE}}$):

$$\mathrm{NAvgSeg}_{\mathrm{SE}} = \frac{\sum \mathrm{Len}(\mathrm{SEG} : \mathrm{SEG}_{\mathrm{SE}})}{N \times \mathrm{Content}_{\mathrm{SEG}_{\mathrm{SE}}}}, \quad (11)$$

where N is the sequence length.

(7) Alternating frequency of α-helices and β-strands and proportion of parallel β-sheets and antiparallel β-sheets (APPA).

Liu and Jia compared the alternating frequencies of different structure elements and found that the α-helices and β-strands alternate more frequently in α/β proteins than in $\alpha + \beta$ proteins, so they introduced the alternating frequency of the α-helices and β-strands to predict protein structural class [35]. The normalized alternating frequency is defined as follows:

$$\mathrm{NAlt}_{\mathrm{SE}} = \frac{\mathrm{Content}_{\alpha-\beta}}{\mathrm{SeqLen}}, \quad (12)$$

where $\text{Content}_{\alpha-\beta}$ denotes the total alternation of the α -helices and β-strands, and SeqLen is the sequence length.

2.6. Recursive Feature Selection with Random Forest. Each decision tree in the random forest is divided into training sets from the root node according to the top-down principle. The root node of the tree is divided into left and right nodes according to the principle of maximum information gain, that is, the training data of the node is divided into two subsets. Under the same rule, the remaining nodes continue to split until the branch stop rule is satisfied. Among them, node information gain can be calculated by information entropy, information gain rate, and Gini index. In this study, information entropy is selected to obtain information gain, which is defined as follows:

$$\text{IG}(S, A) = \text{Entropy}(S) - \text{Entropy}(S, A), \tag{13}$$

where

$$\begin{aligned} \text{Entropy}(S) &= -\sum_{i=1}^{c} p(i)\log_2(p(i)), \\ \text{Entropy}(\text{S}, \text{A}) &= \sum_{v\in\text{values}(A)} \frac{|S_v|}{|S|}\text{Entropy}(S_v), \end{aligned} \tag{14}$$

where S is the training set with the number of categories c, A is the characteristic attribute, $p(i)$ is the probability of the class i in S, $i = 1, \cdots, c$. S_v is the S subset of attribute A, $|S_v|$ is the number of statistical samples, and $|S|$ is the number of samples of training set S. In this study, there are four types of problems; thus, $c = 4$.

For the decision tree classifier, the classification rate is an important index to measure the constructed classifier, but the importance of feature information in the construction of the decision tree node cannot be ignored. In order to select the important features with a strong classification ability, this work introduces the idea of random forest feature selection based on relative importance.

In the experiment, a certain number of features are randomly selected from the candidate features to construct a large number of decision trees, so as to select representative and effective feature information. Firstly, the d candidate features obtained from different feature extraction methods are randomly divided into s subsets. In each subset, 50% of the samples corresponding to m features are randomly selected as the training sample subset, and the remaining 50% as the test sample subset, which are, respectively, used to construct the classification tree and evaluate the performance of the classification tree, t times in total. After the above two steps, a total of st decision trees are generated, in which s and t must be large enough, especially s. Each feature information has a chance to appear in different subsets, and it also makes the selected feature information more accurate.

In order to measure the relative importance of the extracted features, the weighted classification rate is used to evaluate the classification ability of the decision tree on the test set. For a class c classification problem, let n_{ij} be the number of class i samples divided into class j samples, $i, j = 1, \cdots, c$. In this way, the weighted classification rate introduces the size of each class sample set. The specific definition is as follows:

$$w = \frac{1}{c}\sum_{i=1}^{c}\frac{n_{ii}}{n_{i1} + n_{i2} + \cdots + n_{ic}}. \tag{15}$$

In the decision tree, if a feature contains more information, it will play a greater role in the classification rate of the decision tree and gain more information. Therefore, the relative importance (RI) index of a feature is defined as

$$\text{RI}_{g_k} = \sum_{\tau=1}^{st} w \sum_{n_{g_k}} \text{IG}\left(n_{g_k}(\tau)\right)\left(\frac{\text{no.inn}_{g_k}(\tau)}{\text{no.in}\tau}\right), \tag{16}$$

where w is the weighted classification rate of a decision tree. In the st decision trees of random forest, g_k is the relatively important feature generated in the τ tree. All nodes are denoted as $n_{g_k}(\tau)$, $\text{IG}(n_{g_k}(\tau))$ and $\text{no.inn}_{g_k}(\tau)$ are labeled as the information gain and sample number of the nodes, and $\text{no.in}\tau$ is the number of roots of the τ tree. The RI value of each feature is calculated using the above method, and then, the features are sorted according to the RI value. Finally, the representative feature information with great contribution can be selected.

2.7. Classification Algorithm. Support vector machine (SVM) is a large edge classifier based on statistical learning theory. It uses an optimal separation hyperplane to separate two kinds of data. For the binary support vector machine, the decision function is

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b, \tag{17}$$

where b is a constant, C is a cost parameter controlling the trade-off between allowing training errors and forcing rigid margins, $y_i \epsilon \{-1, +1\}$, x_i is the support vector, $0 \le \alpha_i \le C$, and $K(x_i, x)$ is the kernel function. This paper uses Vapnik's support vector machine to predict protein structural classes [41]. Since protein has more than two structural classes, we choose the "one-to-one" strategy of multiclass SVM. Given an unknown class of test protein, we calculate the combined features and select the efficient features based on the recursive feature selection with random forest. The support vector machine will then find an optimized linear partition to solve this multiclass problem.

This work chooses the Gauss kernel function of the support vector machine because of its superiority in solving nonlinear problems [42, 43]. Furthermore, a simple grid search strategy is used to select the parameters C and gamma with the highest overall prediction. It is designed based on 10 times cross-validation of each data set, and the values of C and gamma are taken from the 2^{-10} to 2^{10}.

2.8. Performance Evaluation. There are three widely used cross-validation methods (subsampling test, independent data set test, and jackknife test) to evaluate the classifier's ability. The jackknife test always produces a unique result, which helps to check the quality of various prediction methods. Therefore, we chose the jackknife test to evaluate the proposed method and introduced the sensitivity (Sens), specificity (Spec), and F1 as standard performance indicators, as well as the accuracy and overall accuracy of each category. These standard performance indicators are defined as follows:

$$\begin{aligned} \text{Accuracy}_i &= \frac{\text{TP}_i}{|C_i|}, \\ \text{Overall accuracy} &= \frac{\sum \text{TP}_i}{\sum |C_i|}, \\ \text{Sens} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Spec} &= \frac{\text{TN}}{\text{FP} + \text{TN}}, \\ \text{F1} &= \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}, \end{aligned} \tag{18}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, FN is the number of false negatives, and $|C_i|$ is the number of proteins in each structural class C_i (all-α, all-β, α/β and $\alpha + \beta$ classes).

3. Results and Discussion

3.1. Performance of Proposed Prediction Method. The low sequence homology of 25PDB, D640, FC699, and 1189 was 25%, 25%, 40%, and 40%, respectively. A simple grid search strategy is adopted for C and gamma values based on the 10 times cross-validation of each data set. The sensitivity (Sens), specificity (Spec), and F1 of the proposed method are summarized in Table 2.

Table 2 shows that the prediction performance of the all-α class is the best among the four structural classes, and its sensitivity, specificity, and F1 are higher than 90%. But the lower predictions are related the $\alpha + \beta$ class. From Table 3, we find that the overall accuracy of the method is more than 86% for the four data sets. The overall accuracy of the all-α class was significantly higher than that of other categories, and the accuracy was more than 94%, followed by all categories and categories. It is not difficult to find that the average total accuracy of the $\alpha + \beta$ class of the four data sets is 86.1%, which is 10% lower than that of all classes. These results indicate that it is more difficult to predict the $\alpha + \beta$ class because of the nonnegligible overlap in this category.

3.2. Performance Comparison with the Competing Predictions. This paper further compared the proposed method with the available competing methods. Here, the accuracy of each class and the overall accuracy are chosen as evaluation indexes to evaluate all the prediction methods, and their results are summarized in Table 3. The proposed method is first compared with AADP-PSSM [44], AAC-PSSM-AC [45], and Ding et al.'s method [46] based on the position-specific scoring matrix. Among all the experiments, the proposed method achieves the best performance, with accuracy above 5.4-12.5% better than the next competing Ding et al.'s method [46].

Table 2: Sensitivity (Sens), specificity (Spec), and F1 of the proposed method on four data sets.

Data set	Class	Sens (%)	Spec (%)	F1 (%)
25PDB	All-α	94.81	98.29	95.02
	All-β	95.26	98.13	95.05
	α/β	89.88	95.25	86.39
	$\alpha + \beta$	85.71	97.16	88.52
D640	All-α	97.10	97.81	94.70
	All-β	92.86	99.18	95.02
	α/β	97.18	92.87	90.05
	$\alpha + \beta$	80.70	98.93	87.90
FC699	All-α	97.69	99.45	97.32
	All-β	98.51	99.49	98.70
	α/β	95.23	99.38	97.16
	$\alpha + \beta$	96.34	97.68	88.27
1189	All-α	94.62	96.55	90.95
	All-β	89.80	98.50	92.63
	α/β	82.04	94.20	84.05
	$\alpha + \beta$	81.74	92.95	79.12

As for the 25PDB data set, we further compare the proposed method with the competitive methods: SCPRED [32, 33], MODAS [34], S. Zhang et al. [25], RKS-PPSC [47], Ding et al. [48], Xia et al. [49], L.C. Zhang et al. [36], and S.L. Zhang et al. [16]. It is easy to note that the proposed method achieves the best performance, and the overall accuracy is 91.5%, which is 7.2 percentage points higher than Ding et al.'s method [48]. In D640 data sets, we compare the proposed method with SCEC [38], SCPRED [32, 33], RKS-PPSC [47], Zhang et al. [16], and Kong et al. [20]. The overall accuracy of our method is 91.7%, which is 7-8.1% higher than other competitive methods [16, 20]. As for FC699, the comparison is performed between the proposed method and SCPRED [32, 33], 11 features [35], and Kong et al. [20]. We find that the overall accuracy of this method is 96.7%, which is significantly better than other methods. In the 1189 experiment, SCPRED [32, 33], MODAS [34], RKS-PPSC [47], L.C. Zhang et al. [36], S.L. Zhang et al. [16], and Kong et al. [20] are compared with the proposed method, and we find that the proposed method achieves the best performance among all the competing methods. It is the only prediction method with an overall accuracy of more than 86%, which is 3.1% higher than other competitive methods.

It can be seen from Table 3 that the prediction accuracy of α/β class has been improved. Specifically, the accuracies of the $\alpha + \beta$ class for 25PDB, 1189, 640, and FC699 data sets are 85.7%, 80.7%, 96.3%, and 81.7%, respectively, which are 10.2%, 3.5%, 12.1%, and 8.3% higher than those of the next competitive method, respectively [16, 20]. These results

Table 3: Prediction accuracies (variances in the brackets) of the proposed method for four data sets and comparison with other reported results.

Data set	Method	Prediction accuracy (%)				
		All-α	All-β	α/β	$\alpha+\beta$	Overall
25PDB	AADP-PSSM [44]	69.1	83.7	85.6	35.7	70.7
	AAC-PSSM-AC [45]	85.3	81.7	73.7	55.3	74.1
	SCPRED [32, 33]	92.6	80.1	74.0	71.0	79.7
	MODAS [34]	92.3	83.7	81.2	68.3	81.4
	RKS-PPSC [47]	92.8	83.3	85.8	70.1	82.9
	Ding et al. [46]	95.0	81.3	83.2	77.6	84.3
	Xia et al. [49]	92.6	72.5	71.7	71.0	77.2
	Zhang et al. [36]	95.7	80.8	82.4	75.5	83.7
	Ding et al. [48]	91.7	80.8	79.8	64.0	79.0
	Zhang et al. [16]	94.4	83.3	83.5	73.2	83.6
	This paper	94.8	95.3	89.9	85.7	91.5
D640	SCEC [38]	73.9	61.0	81.9	33.9	62.3
	SCPRED [32, 33]	90.6	81.8	85.9	66.7	80.8
	RKS-PPSC [47]	89.1	85.1	88.1	71.4	83.1
	Ding et al. [46]	92.8	88.3	85.9	66.1	82.7
	Zhang et al. [16]	92.0	81.8	87.6	74.3	83.6
	Kong et al. [20]	94.2	80.5	87.6	77.2	84.5
	This paper	97.1	92.8	97.1	80.7	91.7
FC699	SCPRED [32, 33]	—	—	—	—	87.5
	11 features [35]	97.7	88.0	89.1	84.2	89.6
	Kong et al. [20]	96.2	90.7	96.3	69.5	92.0
	This paper	97.7	98.5	95.2	96.3	96.7
1189	AADP-PSSM [44]	69.1	83.7	85.6	35.7	70.7
	AAC-PSSM-AC [45]	80.7	86.4	81.4	45.2	74.6
	SCPRED [32, 33]	89.1	86.7	89.6	53.8	80.6
	MODAS [34]	92.3	87.1	87.9	65.4	83.5
	RKS-PPSC [47]	89.2	86.7	82.6	65.6	81.3
	Zhang et al. [36]	92.4	84.4	84.4	73.4	83.6
	Ding et al. [46]	89.2	88.8	85.6	58.5	81.2
	Zhang et al. [16]	91.5	86.7	82.0	66.4	81.8
	Kong et al. [20]	91.9	84.4	85.3	72.2	83.5
	This paper	94.6	89.7	82.1	81.7	86.6

show that the proposed method outperforms the available PSSM-based and PSSM-free prediction methods, indicating that the recursive feature selection with the random forest can select important features from the combined feature set and advances predict precision. This understanding can be used to develop more powerful protein structure prediction methods.

3.3. Influence of Recursive Feature Selection. A feature of the proposed method is the recursive feature selection with random forest, which calculates the RI value of each feature and selects the representative features with great contribution. For a better understanding of the recursive feature selection, we select the feature set with size from 10 to 857. All experiments are performed with each selected feature set using the jackknife cross-validation test, and the overall accuracy is chosen to represent the score in this prediction. Figure 1 shows the overall accuracies of all experiments with the selected feature sets for four data sets.

As would be expected, the overall accuracy first increases and then decreases as the selected feature size continues to increase. When the selected feature set size is less than 50, all data sets have reached the best prediction. As the number of selected features increases, the overall accuracy will decrease. The number of selected features corresponding to the best performance is far less than the total number of original features. Therefore, there is a large amount of redundant information in the original combination feature set. After the recursive feature selection with the random forest is used to select and reduce the dimension, the

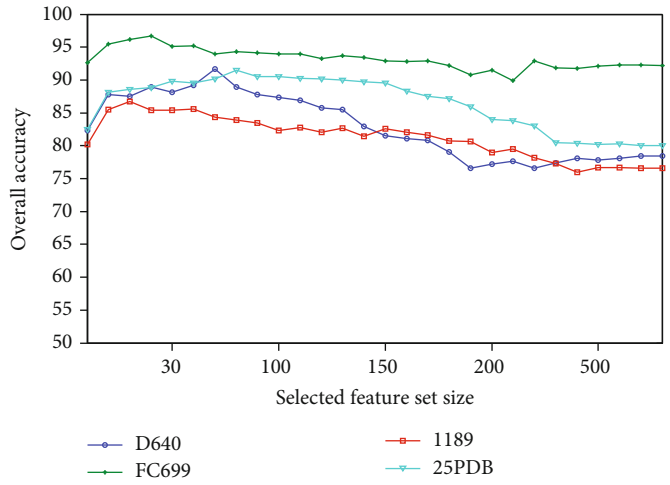

Figure 1: The comparison of the overall accuracies of all experiments with the selected feature sets for four data sets.

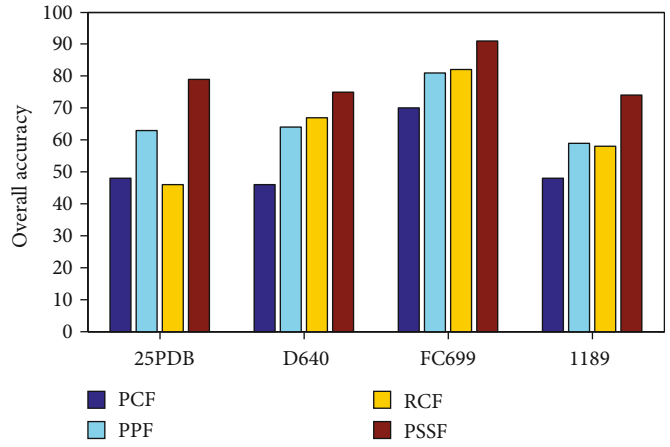

Figure 2: Comparison of the overall prediction accuracies of four kinds of the protein features.

classification rates of four data sets 25PDB, 1189, 640, and FC699 are 91.5%, 86.6%, 91.7%, and 96.7%, respectively, which increased by 4.6-13.3%.

3.4. Influence of the Different Features. To improve the prediction of protein structural classes, we use four kinds of protein features: protein sequence features, protein position features, reduced combined features, and predicted secondary structure features. For brevity, let PSF, PPF, RCF, and PSSF denote these four kinds of protein features, respectively. Through the experiments, we want to address which features contribute to the prediction better.

To evaluate the contribution of each kind of the protein features, we present the comparison of the overall prediction accuracies of four kinds of the protein features in Figure 2. It indicates that each feature makes its own positive contributions to the predictions. PSSF achieves the best performance among the four kinds of the protein features, which is 8%~31% higher than the other three features. In addition, PSSF are selected as the efficient features, which indicates that PSSF is relatively important and has a great contribution to the improvement of prediction. It is easy to note that PSSF is directly extracted from the predicted secondary structure sequences, including the information of α-helix and β-fold alternation frequency and spatial arrangement. Compared with the amino acid frequency and position, the secondary structure sequence information is more closely related to the secondary structure types; this is why it achieves the best performance in protein structure prediction.

4. Conclusion

Protein structural classes provide some useful information for the study of the whole folding type, especially for proteins with low sequence similarity. Various types of protein features are combined to improve the protein structural class prediction. However, it should be noted that the feature fusion will also bring information redundancy and affect the efficiency and accuracy of prediction. This paper proposed a feature selection method for protein structural class prediction, which calculates the RI value of each feature with the random forest and selects the representative features based on each contribution. To do so, we first extracted protein sequence features and protein position features, reduced

combined features, predicted secondary structure features, and used the recursive feature selection with random forest to select the core features for prediction. The experiment results show that the recursive feature selection with the random forest effectively improves the efficiency of protein structural class prediction. Only less than 5% features are used, but the prediction accuracy is improved by 4.6-13.3%. For a better understanding of different protein features, we compared the contribution of each kind of the protein features and found that the predicted secondary structural features achieve the best performance among the four kinds of the protein features, which is 8%~31% higher than the other features. This understanding can be then used to develop more powerful methods for protein structural class prediction.

Data Availability

All the data used to support the findings of this study are available on https://github.com/qidaizstu/recursive-feature-selection.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61772028) and research grants from Zhejiang Provincial Natural Science Foundation of China (LY20F020016).

References

[1] P. Klein and C. Delisi, "Prediction of protein structural class from the amino acid sequence," *Biopolymers*, vol. 25, no. 9, pp. 1659–1672, 1986.

[2] K. C. Chou, "Structural bioinformatics and its impact to biomedical science and drug discovery," *Frontiers in medicinal chemistry*, vol. 3, pp. 455–502, 2006.

[3] M. Levitt and C. Chothia, "Structural patterns in globular proteins," *Nature*, vol. 261, no. 5561, pp. 552–558, 1976.

[4] A. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2004: refinements integrate structure and sequence family data," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 226D–2229, 2004.

[5] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[6] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, and G. Valiente, "Compression-based classification of biological sequences and structures via the universal similarity metric: experimental assessment," *BMC Bioinformatics*, vol. 8, no. 1, p. 252, 2007.

[7] Q. Dai and T. M. Wang, "Comparison study on *k*-word statistical measures for protein: from sequence to 'sequence space'," *BMC Bioinformatics*, vol. 9, no. 1, 2008.

[8] C. Chen, Y. Tian, X. Zou, P. Cai, and J. Mo, "Using pseudo-amino acid composition and support vector machine to predict protein structural class," *Journal of Theoretical Biology*, vol. 243, no. 3, pp. 444–448, 2006.

[9] K. Chou, "Prediction of protein structural classes and subcellular locations," *Current Protein & Peptide Science*, vol. 1, no. 2, pp. 171–208, 2000.

[10] K. D. Kedarisetti, L. A. Kurgan, and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology," *Biochemical and Biophysical Research Communications*, vol. 348, no. 3, pp. 981–988, 2006.

[11] Q. Dai, L. Wu, and L. H. Li, "Improving protein structural class prediction using novel combined sequence information and predicted secondary structural features," *Journal of Computational Chemistry*, vol. 32, no. 16, pp. 3393–3393, 2011.

[12] K. C. Chou, "A key driving force in determination of protein structural classes," *Biochemical and Biophysical Research Communications*, vol. 264, no. 1, pp. 216–224, 1999.

[13] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[14] R. Y. Luo, Z. P. Feng, and J. K. Liu, "Prediction of protein structural class by amino acid and polypeptide composition," *European Journal of Biochemistry*, vol. 269, no. 17, pp. 4219–4225, 2002.

[15] X. D. Sun and R. B. Huang, "Prediction of protein structural classes using support vector machines," *Amino Acids*, vol. 30, no. 4, pp. 469–475, 2006.

[16] S. L. Zhang, Y. Y. Liang, and X. G. Yuan, "Improving the prediction accuracy of protein structural class: approached with alternating word frequency and normalized Lempel-Ziv complexity," *Journal of Theoretical Biology*, vol. 341, pp. 71–77, 2014.

[17] Y. S. Ding, T. L. Zhang, and K. C. Chou, "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network," *Protein and peptide letters*, vol. 14, no. 8, pp. 811–815, 2007.

[18] L. Wu, Q. Dai, B. Han, L. Zhu, and L. H. Li, "Combining sequence information and predicted secondary structural feature to predict protein structural classes," in *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, pp. 1–4, 2011.

[19] B. Liao, Q. Xiang, and D. Li, "Incorporating secondary features into the general form of Chou's PseAAC for predicting protein structural class," *Protein and peptide letters*, vol. 19, no. 11, pp. 1133–1138, 2012.

[20] L. Kong, L. C. Zhang, and J. F. Lv, "Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 344, pp. 12–18, 2014.

[21] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 452, pp. 22–34, 2018.

[22] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, and L. Yang, "PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition," *Bioinformatics*, vol. 33, no. 1, pp. 122–124, 2017.

[23] K. C. Chou and Y. D. Cai, "Predicting protein structural class by functional domain composition," *Biochemical and Biophysical Research Communications*, vol. 321, no. 4, pp. 1007–1009, 2004.

[24] L. Kurgan, K. Cios, and K. Chen, "SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences," *BMC Bioinformatics*, vol. 9, pp. 1–15, 2008.

[25] S. Zhang, S. Ding, and T. Wang, "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure," *Biochimie*, vol. 93, no. 4, pp. 710–714, 2011.

[26] Q. Dai, Y. Li, X. Liu et al., "Comparison study on statistical features of predicted secondary structures for protein structural class prediction: from content to position," *BMC Bioinformatics*, vol. 14, no. 1, p. 152, 2013.

[27] H. Ding, H. Lin, W. Chen et al., "Prediction of protein structural classes based on feature selection technique," *Interdisciplinary Sciences*, vol. 6, pp. 235–240, 2014.

[28] C. Chen, L. X. Chen, X. Y. Zou, and P. X. Cai, "Predicting protein structural class based on multi-features fusion," *Journal of Theoretical Biology*, vol. 253, no. 2, pp. 388–392, 2008.

[29] A. V. Kumar, R. F. M. Ali, C. Yu, and V. V. Krishnan, "Application of data mining tools for classification of protein structural class from residue based averaged NMR chemical shifts," *Biochimica et Biophysica Acta*, vol. 1854, pp. 1545–1552, 2015.

[30] L. Nanni, S. Brahnam, and A. Lumini, "Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 360, pp. 109–116, 2014.

[31] J. Wang, C. Wang, J. Cao, X. Liu, Y. Yao, and Q. Dai, "Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features," *Gene*, vol. 554, no. 2, pp. 241–248, 2015.

[32] L. A. Kurgan and L. Homaeian, "Prediction of structural classes for protein sequences and domains–impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy," *Pattern Recognition*, vol. 39, no. 12, pp. 2323–2343, 2006.

[33] C. Zheng and L. Kurgan, "Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments," *BMC Bioinformatics*, vol. 9, no. 1, pp. 430–430, 2008.

[34] M. J. Mizianty and L. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," *BMC Bioinformatics*, vol. 10, no. 1, pp. 414–414, 2009.

[35] T. Liu and C. Z. Jia, "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information," *Journal of Theoretical Biology*, vol. 267, no. 3, pp. 272–275, 2010.

[36] L. C. Zhang, X. Q. Zhao, and L. Kong, "A protein structural class prediction method based on novel features," *Biochimie*, vol. 95, no. 9, pp. 1741–1744, 2013.

[37] L. Kurgan and K. Chen, "Prediction of protein structural class for the twilight zone sequences," *Biochemical and Biophysical Research Communications*, vol. 357, no. 2, pp. 453–460, 2007.

[38] K. Chen, L. A. Kurgan, and J. S. Ruan, "Prediction of protein structural class using novel evolutionary collocation-based sequence representation," *Journal of Computational Chemistry*, vol. 29, no. 10, pp. 1596–1604, 2008.

[39] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices[1]," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999.

[40] H. Shen and K. C. Chou, "Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM," *Protein Engineering Design and Selection*, vol. 20, no. 11, pp. 561–567, 2007.

[41] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 2000.

[42] T. Li, K. Fan, J. Wang, and W. Wang, "Reduction of protein sequence complexity by residue grouping," *Protein Engineering*, vol. 1, pp. 323–330, 2003.

[43] Y. Cai, X. Liu, X. Xu, and K. Chou, "Prediction of protein structural classes by support vector machines," *Computers & Chemistry*, vol. 26, no. 3, pp. 293–296, 2002.

[44] T. Liu, X. Zheng, and J. Wang, "Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile," *Biochimie*, vol. 92, no. 10, pp. 1330–1334, 2010.

[45] T. Liu, X. Geng, X. Zheng, R. Li, and J. Wang, "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles," *Amino Acids*, vol. 42, no. 6, pp. 2243–2249, 2012.

[46] S. Y. Ding, S. J. Yan, S. Qi, Y. Li, and Y. H. Yao, "A protein structural classes prediction method based on PSI-BLAST profile," *Journal of Theoretical Biology*, vol. 353, pp. 19–23, 2014.

[47] J. Y. Yang, Z. L. Peng, and X. Chen, "Prediction of protein structural classes for low-homology sequences based on predicted secondary structure," *BMC Bioinformatics*, vol. 11, no. S1, 2010.

[48] S. Y. Ding, S. L. Zhang, Y. Li, and T. M. Wang, "A novel protein structural classes prediction method based on predicted secondary structure," *Biochimie*, vol. 94, no. 5, pp. 1166–1171, 2012.

[49] X. Y. Xia, M. Ge, Z. X. Wang, and X. M. Pan, "Accurate prediction of protein structural class," *PLoS One*, vol. 7, no. 6, p. e37653, 2012.